DOCUMENT RESUME

ED 368 773 TM 021 202

AUTHOR Brennan, Robert L.
TITLE Some Measurement Characteristics of Aggregated versus Individual Scores.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa.
REPORT NO ACT-RR-93-10
PUB DATE Dec 93
NOTE 23p.
AVAILABLE FROM ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243; American College Testing, 2201 North Dodge Street, Iowa City, IA 52243.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Statistical Data (110)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Course Evaluation; *Generalizability Theory; *Measurement Techniques; *Reliability; Sampling; Scores; *Statistical Studies
IDENTIFIERS *Aggregation (Data); American College Testing Program; Performance Based Evaluation; *Variance (Statistical)

ABSTRACT

Not infrequently, investigators assume that reliability for groups is greater than reliability for persons, or that the error variance for groups is less than that for persons. Using generalizability theory, it is shown that this "conventional wisdom" is not necessarily true. Examples are provided from the course-evaluation and the performance-testing literature. In the cases considered in this paper, the conventional wisdom necessarily holds only for comparative statements about person versus group error variance when the universe of generalization has persons fixed and items random. In all other cases, the conventional wisdom may be false, in particular when the generalization is over both samples of persons and samples of items, which often represents the most sensible universe of generalization. An appendix elaborates on reliability. (Contains 8 references.) (Author/SLD)

TM

ACT Research Report Series 93-10

Some Measurement Characteristics of Aggregated Versus Individual Scores

Robert L. Brennan

December 1993

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

Some Measurement Characteristics of Aggregated Versus Individual Scores

Robert L. Brennan

American College Testing

Abstract

Not infrequently, investigators assume that reliability for groups is greater than reliability for persons, and/or error variance for groups is less than error variance for persons. Using generalizability theory, it is shown that this "conventional wisdom" is not necessarily true. Examples are provided from the course evaluation literature and the performance testing literature.

Some Measurement Characteristics of Aggregated Versus Individual Scores

It is often stated that if a test is not reliable enough for making decisions about individuals, or if error variance for individuals is unacceptably large, then the test should be used only for making decisions about groups. Implicit in such statements is an assumption that reliability for groups is necessarily larger than reliability for persons, and error variance for groups is necessarily smaller than error variance for persons. In this paper, such statements or assumptions will be called the "conventional wisdom." The purpose of this paper is to show that this conventional wisdom is not necessarily true, and to identify specific conditions that lead to contradictions of this conventional wisdom.

These issues are considered in the context of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), without a full explication of all the details of the theory. Readers unfamiliar with generalizability theory can consult Cronbach et al. (1972), Brennan (1992a), or Shavelson and Webb (1991). Also, many aspects of reliability (or generalizability) of group means have been treated by Kane and Brennan (1977). A brief introduction to generalizability theory is provided by Bennan (1992b).

Generalizability Coefficients for Groups with Two Random Facets

When persons (p) are nested within groups (g) and crossed with items (i), the design is denoted ($p:g$) x i, and the linear model is

$$X_{pi:g} = \mu + \mu_g{\sim} + \mu_{p:g}{\sim} + \mu_i{\sim} + \mu_{gi}{\sim} + \mu_{pi:g}{\sim} . \qquad (1)$$

The terms to the right of the equality (except μ) are uncorrelated score effects with expectations of zero, and the $\mu_{pi:g}{\sim}$ term is the interaction effect confounded with other sources of error. The variances of these score effects are called variance components.

For this design, if groups are the objects of measurement, then the universe of generalization consists of the p and i facets. If, in addition, p and i are both random, then the generalizability coefficient for generalizing over samples of k items and n persons within each group is

$$E\rho_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{p:g}^2/n + \sigma_{gI}^2 + \sigma_{pI:g}^2/n} \quad , \tag{2}$$

where $\sigma_{gI}^2 = \sigma_{gi}^2/k$ and $\sigma_{pI:g}^2 = \sigma_{pi:g}^2/k$.

It is important to note that n in Equation 2 is the number of persons within a group, not the larger number of persons across all groups.

Assuming that p and i are both random implies that replications of the measurement procedure would involve different sets of persons and items or, equivalently, that an investigator wants to generalize to a larger set of persons and items than those in a particular measurement procedure. This assumption seems especially sensible for programs such as NAEP, which uses a type of matrix sampling design.

From the perspective of generalizability theory, traditional measurement error is most closely associated with generalizing over samples of items. It is important to note, however, that traditional measurement error is not necessarily the only, or even the most important, source of unreliability for inferences about group means. As noted by Feldt and Brennan (1989):

> The test results for any given year reflect not only the character of the instructional program but also the character of students enrolled at that specific moment. These individuals must be regarded as a sample, in a longitudinal sense, from the population that flows through the district schools over a period of years. A curricular judgment can be in error if a particular year's class happens to be unusually strong or weak. Thus, even

> if authorities were privileged to know the true scores of current students, there could be substantial sampling error in using the results of one class to infer something about the impact of a program. An estimate of the reliability of class means must take this into account. (p. 127)

In short, in most circumstances, generalizing over both items and persons seems sensible in examining the reliability of group means.

Using the notational system introduced above, when persons within a single randomly selected group are the objects of measurement, the generalizability coefficient is

$$E\rho^2_{p:g} = \frac{\sigma^2_{p:g}}{\sigma^2_{p:g} + \sigma^2_{pI:g}} \quad . \tag{3}$$

It is important to note that Equation 3 is for persons within a group, not across groups. The generalizability coefficient for all persons across groups is

$$E\rho^2_{p} = \frac{\sigma^2_{g} + \sigma^2_{p:g}}{\sigma^2_{g} + \sigma^2_{p:g} + \sigma^2_{gI} + \sigma^2_{pI:g}} \quad . \tag{4}$$

Note that for both of the coefficients in Equations 3 and 4, persons are the objects of measurement and the universe of generalization involves the items facet, only.

Usually, when comparative statements are made about reliability coefficients for groups and persons, the intended interpretation of reliability for persons is given by Equation 4. Therefore, a central focus of this paper is to compare Equation 2 and Equation 4. In particular, it is of interest to identify conditions under which $E\rho^2_g < E\rho^2_p$. One such condition is $\sigma^2_g = 0$, which is an unlikely occurrence implying that all group means are equal.

The inequality $E\rho^2_g < E\rho^2_p$ is also true when $k \to \infty$ because in that case $E\rho^2_p = 1$ and $E\rho^2_g = \sigma^2_g/(\sigma^2_g + \sigma^2_{p:g}/n) < 1$. Consequently, it seems likely that long tests that are

highly reliable for decisions about persons will be less reliable for decisions about groups. For example, the following estimated variance components were obtained from an administration of the ACT Assessment Mathematics test in schools (i.e., groups) in a particular state:

$$\hat{\sigma}^2_g = .0016,\ \hat{\sigma}^2_{p:g} = .0329,\ \hat{\sigma}^2_{gi} = .0009,\ \text{and}\ \hat{\sigma}^2_{pi:g} = .1809.$$

The ACT Mathematics test contains $k = 60$ multiple choice items, and the average number of students per school was about $n = 145$. For these values

$$E\hat{\rho}^2_o = .86 < E\hat{\rho}^2_p = .92\,.$$

Even with such a large value for n, $E\rho^2_g$ is still less than $E\rho^2_p$ in large part because the ACT Mathematics test is very reliable for person-level decisions.

Equations 2 and 4 also imply that $E\rho^2_g < E\rho^2_p$ is true when $n = 1$, which suggests that $E\rho^2_g < E\rho^2_p$ is more likely to be true for small values of n than for large values. However, in general, for $1 < n < \infty$ and $1 < k < \infty$ there appears to be no simple, necessary relationship among the variance components that guarantees that $E\rho^2_g < E\rho^2_p$. Even so, there are sufficient conditions that do pertain. One such condition is discussed next.

A Sufficient Condition

Note that, for a given value of k, the maximum value of Equation 2 occurs when $n \to \infty$, in which case

$$(E\rho^2_g \mid n \to \infty) = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_{gI}}\quad . \tag{5}$$

Clearly, if $(E\rho_g^2 \mid n \to \infty) < E\rho_p^2$ then $E\rho_g^2 < E\rho_p^2$. Therefore, whenever Equation 5 is smaller than Equation 4, $E\rho_g^2 < E\rho_p^2$. Consequently, we will focus on Equations 4 and 5 to obtain a sufficient condition for $E\rho_g^2$ to be smaller than $E\rho_p^2$. Letting

$$K = \sigma_g^2/\sigma_{p:g}^2 \quad \text{and} \tag{6}$$

$$L = \sigma_{gI}^2/\sigma_{pI:g}^2 \ , \tag{7}$$

Equation 4 can be written

$$E\rho_p^2 = \frac{\sigma_g^2 + \sigma_g^2/K}{\sigma_g^2 + \sigma_g^2/K + \sigma_{gI}^2 + \sigma_{gI}^2/L}$$

$$= \frac{\left(\frac{K+1}{K}\right)\sigma_g^2}{\left(\frac{K+1}{K}\right)\sigma_g^2 + \left(\frac{L+1}{L}\right)\sigma_{gI}^2}$$

$$= \frac{\sigma_g^2}{\sigma_g^2 + \left(\frac{K}{K+1}\right)\left(\frac{L+1}{L}\right)\sigma_{gI}^2} \ . \tag{8}$$

Equation 5 is smaller than Equation 8 when

$$\left(\frac{K}{K+1}\right)\left(\frac{L+1}{L}\right) < 1 \ ,$$

which implies that $(KL + K) < (KL + L)$. This inequality is true whenever $K < L$.

It follows that, $E\rho_g^2 < E\rho_p^2$

whenever

$$\sigma_g^2/\sigma_{p:g}^2 \ < \ \sigma_{gI}^2/\sigma_{pI:g}^2 \ , \tag{9}$$

or equivalently whenever

$$\frac{\sigma_g^2}{\sigma_g^2 + \sigma_{p:g}^2} < \frac{\sigma_{gI}^2}{\sigma_{gI}^2 + \sigma_{pI:g}^2} \quad . \tag{10}$$

Loosely speaking, these results mean that reliability for groups is less than reliability for persons whenever the proportion of persons' universe score variance attributable to groups is less than the proportion of persons' error variance attributable to groups. (This statement is "loose" primarily because it does not explicitly specify that generalization is over both persons and items from the infinite universe of generalization for both facets.) The condition in Equation 9 or Equation 10 might hold, for example, if schools (i.e., groups) have highly similar universe scores, but at a particular time students in different schools have been exposed to different subsets of the tested topics.

Since $\sigma_{gI}^2 = \sigma_{gi}^2/k$ and $\sigma_{pI:g}^2 = \sigma_{pi:g}^2/k$, the right side of Inequality 10 is invariant over k . Consequently, this inequality is equivalent to

$$\frac{\sigma_g^2}{\sigma_g^2 + \sigma_{p:g}^2} < \frac{\sigma_{gi}^2}{\sigma_{gi}^2 + \sigma_{pi:g}^2} \quad , \tag{11}$$

which is sometimes more convenient to use. In short, $E\rho_g^2 < E\rho_p^2$ if Inequality 9, 10, or 11 holds. As noted previously, this is a sufficient condition for $E\rho_g^2 < E\rho_p^2$ -- it is by no means a necessary condition. That is, $E\rho_g^2$ can be smaller than $E\rho_p^2$ even if Inequality 9, 10, or 11 does not hold.

Two Examples

Discussed next are two examples that illustrate circumstances under which $E\hat{\rho}_g^2 < E\hat{\rho}_p^2$. The first example is from the course evaluation literature. It illustrates a circumstance under which the sufficient condition in Equations 9, 10, or 11 is satisfied. The second example is from the performance testing literature.

Example 1. Kane, Gillmore, and Crooks (1976) studied the generalizability of class (i.e., group) means in the context of student evaluations of teaching. One of the questionnaires they studied was administered in all courses taught in the Physics

Department at the University of Illinois, Urbana-Champaign. "Fifteen classes that had twenty or more students were randomly selected, with the restriction that only one section taught by each instructor was included in the sample (Kane et al., 1976, p. 177)." Thus, there is a linking of each class with a unique instructor, and generalizations about class means are effectively generalizations about instructors.

The questionnaire contained a set of $k = 8$ "attribute" items (e.g., ability to answer questions) that were analyzed separately from other items. For these items,

$$\hat{\sigma}^2_g = .03,\ \hat{\sigma}^2_{p:g} = .17,\ \hat{\sigma}^2_{gi} = .05, \text{ and } \hat{\sigma}^2_{pi:g} = .28 .$$

Using Equation 11

$$\frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \hat{\sigma}^2_{p:g}} = \frac{.03}{.20} = .150 < \frac{\hat{\sigma}^2_{gi}}{\hat{\sigma}^2_{gi} + \hat{\sigma}^2_{pi:g}} = \frac{.05}{.33} = .152 .$$

Therefore, these data satisfy a sufficient condition for $E\hat{\rho}^2_g < E\hat{\rho}^2_p$, which means that $E\hat{\rho}^2_g < E\hat{\rho}^2_p$ for <u>all</u> pairs of values for n and k. For example, if $n = 20$ and $k = 8$,

$$E\hat{\rho}^2_g = .65 < E\hat{\rho}^2_p = .83 .$$

<u>Example 2</u>. Shavelson, Baxter, and Gao (1993) provide an extensive discussion of the sampling variability of performance assessments in the context of several data sets. One such data set is from the California Assessment Program (CAP) which conducted a voluntary statewide science assessment in 1989-1990 with approximately 600 schools. Students took five performance tasks involving identifying materials that serve as conductors, classifying leaves, identifying unknown rocks, estimating and measuring characteristics of water, and discovering reasons why fish were dying. The design employed by Shavelson et al. (1993) is mc e complicated than the $(p:g) \times i$ design considered in this paper, but a subset of their results gives

$\hat{\sigma}^2_g = .07$, $\hat{\sigma}^2_{p:g} = .23$, $\hat{\sigma}^2_{gi} = .07$, and $\hat{\sigma}^2_{pi:g} = .43$.

In this case, $\hat{\sigma}^2_g/(\hat{\sigma}^2_g + \hat{\sigma}^2_{p:g}) = .23$, $\hat{\sigma}^2_{gi}/(\hat{\sigma}^2_{gi} + \hat{\sigma}^2_{pi:g}) = .14$, and Inequality 11 is not satisfied. However, there are numerous combinations of values of n and k for which $E\hat{\rho}^2_g < E\hat{\rho}^2_p$. For example, when $n = 20$ and $k = 5$

$$E\hat{\rho}^2_g = .70 < E\hat{\rho}^2_p = .75 .$$

Recalling that the CAP involved $k = 5$ performance tasks, it can be shown that $E\hat{\rho}^2_g < E\hat{\rho}^2_p$ whenever $n \leq 33$. Furthermore, for any value of k, $E\hat{\rho}^2_g < E\hat{\rho}^2_p$ when $n_p \leq 14$. In other words, no matter how many performance tasks are employed in the CAP, the conventional wisdom about group reliability being larger than person reliability will be incorrect if the number of persons within schools is less than 15.

These examples illustrate that $E\rho^2_g < E\rho^2_p$ is not simply a mathematical possibility. It is a likely occurrence in numerous circumstances, especially when variability of persons within groups, $\sigma^2_{p:g}$, is relatively large.

Error Variances for Groups With Two Random Facets

Just as the conventional wisdom suggests that reliability for groups is greater than reliability for persons, so too investigators often implicitly or explicitly assume that error variance for groups is less than error variance for persons. However, as discussed next, this conventional wisdom about error variances is not necessarily true, either.

The error variance associated with $E\rho^2_g$ in Equation 2 is the relative error variance for group mean scores when generalization is over both persons and items:

$$\sigma^2(\delta_g) = \sigma^2_{p:g}/n + \sigma^2_{gI} + \sigma^2_{pI:g}/n . \qquad (12)$$

For $E\rho_p^2$ in Equation 4 the relative error variance for person mean scores when generalization is over items is

$$\sigma^2(\delta_p) = \sigma^2_{gI} + \sigma^2_{pI:g} \quad . \tag{13}$$

We wish to determine conditions under which $\sigma^2(\delta_g) > \sigma^2(\delta_p)$.

From Equations 12 and 13, $\sigma^2(\delta_g) > \sigma^2(\delta_p)$ whenever

$$\sigma^2_{p:g}/n + \sigma^2_{pI:g}/n > \sigma^2_{pI:g} .$$

Multiplying both sides by n and collecting terms gives

$$\sigma^2_{p:g} > (n - 1)\, \sigma^2_{pI:g} \quad .$$

Dividing both sides by $\sigma^2_{pI:g}$ gives

$$\sigma^2_{p:g}/\sigma^2_{pI:g} > (n - 1) \tag{14}$$

The left side of the above inequality is the signal/noise ratio associated with the generalizability coefficient in Equation 3 for the reliability of persons within a randomly selected group, $E\rho^2_{p:g}$. Since a generalizability coefficient can be viewed as the ratio of signal (i.e., universe score variance) to signal plus noise (i.e., relative error variance), Inequality 14 is equivalent to

$$\frac{\sigma^2_{p:g}}{\sigma^2_{p:g} + \sigma^2_{pI:g}} > \frac{n - 1}{(n - 1) + 1} \quad .$$

Therefore, a necessary condition for $\sigma^2(\delta_g) > \sigma^2(\delta_p)$ is

$$E\rho^2_{p:g} > (n - 1)/n \quad . \tag{15}$$

[These results for relative error variance, $\sigma^2(\delta)$, also hold for absolute error variance, $\sigma^2(\Delta)$, because for both groups and persons $\sigma^2(\Delta)$ differs from $\sigma^2(\delta)$ by the same constant, σ^2_i/k -- i.e. $\sigma^2(\Delta_g) - \sigma^2(\delta_g) = \sigma^2(\Delta_p) - \sigma^2(\delta_p) = \sigma^2_i/k$.]

Clearly, when $n \rightarrow \infty$ Inequality 15 will not hold. So, for large values of n, it is reasonable to assume that error variance for persons is likely to be larger than error variance for groups. However, for small values of n, this need not be so. For example, using Inequality 15, $\sigma^2(\delta_g) > \sigma^2(\delta_p)$ in the following cases:

$n < 20$ and $E\rho^2_{p:g} = .95$, and

$n < 10$ and $E\rho^2_{p:g} = .90$.

These cases are not so extreme as to be entirely implausible, especially for long tests. Consequently, it is unwise to assume that error variance for person mean scores is always greater than error variance for group mean scores.

Note that, if $\sigma^2(\delta_g) > \sigma^2(\delta_p)$ it necessarily follows that $E\rho^2_g < E\rho^2_p$. To see this, recall from Equations 2 and 4 that universe score variance for groups can not be larger than universe score variance for persons. This guarantees that $E\rho^2_g < E\rho^2_p$ when $\sigma^2(\delta_g) > \sigma^2(\delta_p)$. In other words, Equation 15 is another sufficient condition for $E\rho^2_g < E\rho^2_p$.

By contrast, if $E\rho^2_g < E\rho^2_p$ it does not necessarily follow that $\sigma^2(\delta_g) > \sigma^2(\delta_p)$. This is true for the Shavelson et al. (1993) example with $n = 20$ and $k = 5$. In this case, it has already been shown that $E\hat{\rho}^2_g < E\hat{\rho}^2_p$, and $\hat{\sigma}^2(\delta_g) = .03 < \hat{\sigma}^2(\delta_p) = .10$. Therefore, it is possible for group error variance to be smaller than person error variance (in accord with the conventional wisdom), while at the same time group reliability is less than person reliability (against the conventional wisdom.)

One Random Facet

To this point it has been assumed that, when group means are the objects of measurement, both persons and items are randomly sampled from an infinite universe of generalization. Call this the "unrestricted" universe. It has been argued in the introduction to the previous section, that in most circumstances this unrestricted universe is sensible for examining the measurement charactristics of aggregated scores. However, there may be circumstances in which it is reasonable to consider a restricted universe of generalization in which persons are fixed and items are random. If so, then replications of the measurement procedure would involve the same persons but different sets of items. For this restricted universe, generalizability coefficients will be larger, and error variances will be smaller, than for the unrestricted universe. Consequently, it is more likely that for this restricted universe the conventional wisdom about group reliability and error variance holds.

When persons are fixed and items are random, the generalizability coefficient for groups is

$$E\rho_g^2 \mid P = \frac{\sigma_g^2 + \sigma_{p:g}^2/n}{\sigma_g^2 + \sigma_{p:g}^2/n + \sigma_{gI}^2 + \sigma_{pI:g}^2/n} , \qquad (16)$$

and the associated relative error variance for group means is

$$\sigma^2(\delta_g \mid P) = \sigma_{gI}^2 + \sigma_{pI:g}^2/n . \qquad (17)$$

In effect, fixing persons causes $\sigma_{p:g}^2/n$ to move from relative error variance (see Equation 12) to universe score variance.

Comparing Equation 17 with Equation 13 shows that when generalization is over items only, $\sigma^2(\delta_g \mid P) < \sigma^2(\delta_p)$ as long as $n > 1$. That is, the conventional wisdom about error variances holds. However, the conventional wisdom about reliability coefficients

does not necessarily hold. In particular, as shown in the appendix, $(E\rho_g^2 \mid P) < E\rho_p^2$ when Inequality 9, 10, or 11 holds. That is, if Inequality 9, 10, or 11 holds, then it necessarily follows that $(E\rho_g^2 \mid P) < E\rho_p^2$ in the restricted universe. By contrast, Inequality 9, 10, or 11 is only a sufficient condition for $E\rho_g^2 < E\rho_p^2$ in the unrestricted universe.

It is also possible to consider $E\rho_g^2 \mid I$ and $\sigma^2(\delta_g \mid I)$ for the case when items are fixed and persons are random. If so, then replications of the measurement procedure would involve the same items but different sets of persons. This possibility is considered by Feldt and Brennan (1989, pp. 127, 135-136). It can be shown that there are conditions such that $(E\rho_g^2 \mid I) < E\rho_p^2$ and $\sigma^2(\delta_g \mid I) > \sigma^2(\delta_p)$. However, there is a conceptual conflict in comparing the magnitude of $E\rho_p^2$ with $E\rho_g^2 \mid I$, and the magnitude of $\sigma^2(\delta_p)$ with $\sigma^2(\delta_g \mid I)$. The conflict arises because items are fixed for $E\rho_g^2 \mid I$ and $\sigma^2(\delta_g \mid I)$, whereas items are random for $E\rho_p^2$ and $\sigma^2(\delta_p)$. Therefore, although statements can be made about the relative magnitudes of these quantities, such comparisons are likely to be misleading.

Summary and Discussion

It has been shown that when persons and items are random the conventional wisdom that $E\rho_g^2 > E\rho_p^2$ and $\sigma^2(\delta_g) < \sigma^2(\delta_p)$ does not necessarily hold. In particular, $\sigma_g^2/(\sigma_g^2 + \sigma_{p:g}^2) < \sigma_{gi}^2/(\sigma_{gi}^2 + \sigma_{pi:g}^2)$ is a sufficient condition for $E\rho_g^2 < E\rho_p^2$. Even if this sufficient condition is not met, there can be various combinations of values for n and k such that $E\rho_g^2 < E\rho_p^2$. Also, contrary to the conventional wisdom, $\sigma^2(\delta_g) > \sigma^2(\delta_p)$ whenever $E\rho_{p:g}^2 > (n - 1)/n$, and for small values of n and long tests this condition might well be met.

When persons are fixed and items are random, the conventional wisdom that $\sigma^2(\delta_g \mid P) < \sigma^2(\delta_p)$ is true provided $n > 1$. However, for this restricted universe of generalization, the conventional wisdom that $(E\rho_g^2 \mid P) > E\rho_p^2$ is false if $\sigma_g^2/(\sigma_g^2 + \sigma_{p:g}^2) < \sigma_{gi}^2/(\sigma_{gi}^2 + \sigma_{pi:g}^2)$.

In short, for the cases considered in this paper, the conventional wisdom necessarily holds only for comparative statements about person vs. group error variance

when the universe of generalization has persons fixed and items random. In all other cases, the conventional wisdom about reliability coefficients and error variances may be false. In particular, the conventional wisdom may be false when generalization is over both samples of persons and samples of items, which often represents the most sensible universe of generalization. For this universe, the form of Equations 2, 4, 12, and 13 clearly shows that $\sigma^2_{p:g}$ is incorporated in universe score variance when persons are the objects of measurement, whereas $\sigma^2_{p:g}$ is incorporated in error variance when groups are the object of measurement. Therefore, the magnitude of $\sigma^2_{p:g}$ is likely to be very influential in whether or not the conventional wisdom holds.

As illustrated by the examples in this paper, violations of the conventional wisdom about group means are not merely mathematical possibilities -- such violations are quite common, although they are seldom reported.

Some of the results presented in this paper may seem to contradict the central limit theorem. In its simplest form, the central limit theorem implies that error variance for mean scores will be less than error variance for individual scores. This has been shown to be true for a universe of generalization in which items constitute the only random facet, but not necessarily true for a universe of generalization in which both persons and items are random. One of the strengths of generalizability theory is that it permits an investigator to disentangle the amount of error attributable to multiple facets. This cannot be done (or at least not directly) using the simple form of the central limit theorem.

Sometimes investigators appear to assume that statements about the relative magnitudes of reliability coefficients are interchangeable with statements about the relative magnitudes of the corresponding error variances. As discussed previously, this is not necessarily true. It is possible that $E\rho^2_g < E\rho^2_p$ (against the conventional wisdom) while $\sigma^2(\delta_g) < \sigma^2(\delta_p)$ (in accord with the conventional wisdom). It is important,

therefore, that investigators not generalize from statements about reliability to statements about error variance, or vice-versa.

Throughout this paper, emphasis has been on discussing conditions under which $E\rho_g^2 < E\rho_p^2$ and $\sigma^2(\delta_g) > \sigma^2(\delta_p)$ -- i.e., conditions under which the conventional wisdom is reversed. Note, as well, that even if a reversal does not occur, aggregation to a group level may have relatively little impact on reliability. For example, for the Shavelson et al. (1993) example introduced earlier, if $k = 5$, there is no value of n such that $E\hat{\rho}_g^2$ is greater than $E\hat{\rho}_p^2$ by .10 or more, and $n > 90$ is required for $E\hat{\rho}_g^2$ to be greater than $E\hat{\rho}_p^2$ by .05. Furthermore, especially for relatively small values of k, $E\hat{\rho}_g^2$ may be unacceptably small even if it is larger than $E\hat{\rho}_p^2$. This is a distinct possibility in some performance testing contexts.

Many writers, including this author, have argued that too frequently reliability coefficients are referenced in contexts where error variances would be more appropriate Also, in item response theory, there is little attention given to reliability coefficients. These two points may seem to suggest that the issues raised in this paper about reliability coefficients do not deserve much attention. Such a conclusion would be unfortunate. Reliability coefficients (as well as signal/noise ratios) have the distinct advantage of providing, in one statistic, a comparison between true (or universe) score variance and error variance, whereas examining error variance in isolation often leaves an investigator pondering whether or not error variance is to be considered large or small. This is a particularly important consideration when examining group means. Aggregation may well lead to a sizable decrease in error variance, but this can be very misleading if an investigator fails to take into account the corresponding decrease in true (or universe) score variance. In short, both reliability coefficients and error variances have utility for examining the measurement characteristics of aggregated scores versus individual scores.

References

Brennan, R. L. (1992a). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: American College Testing.

Brennan, R. L. (1992b). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27-34.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), 105-146. New York: Macmillan.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, *47*, 267-292.

Kane, M. T., Gillmore, G. M., & Crooks, T. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, *13*, 171-183.

Shavelson, R. J., Baxter, G. P., & Gao, Xiaohong. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215-232.

Shavelson, R J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park CA: Sage.

Author Note

The author gratefully acknowledges that Dean Colton, Michael Kane, and Xiaohong Gao made helpful comments and suggestions on a previous draft of this paper.

Appendix

A Condition Under Which $(E\rho_g^2 \mid P) < E\rho_p^2$

Let $K = \sigma_g^2/\sigma_{p:g}^2$ and $L = \sigma_{gI}^2/\sigma_{pI:g}^2$ as in Equations 6 and 7, respectively. It follows from Equation 8 that

$$E\rho_p^2 = \frac{\sigma_g^2}{\sigma_g^2 + \left(\frac{K}{L}\right)\left(\frac{L+1}{K+1}\right)\sigma_{gI}^2} \quad . \tag{A1}$$

From Equation 16

$$E\rho_g^2 \mid P = \frac{\sigma_g^2 + \sigma_g^2/Kn}{\sigma_g^2 + \sigma_g^2/Kn + \sigma_{gI}^2 + \sigma_{gI}^2/Ln}$$

$$= \frac{\left(\frac{Kn+1}{Kn}\right)\sigma_g^2}{\left(\frac{Kn+1}{Kn}\right)\sigma_g^2 + \left(\frac{Ln+1}{Ln}\right)\sigma_{gI}^2}$$

$$= \frac{\sigma_g^2}{\sigma_g^2 + \left(\frac{K}{L}\right)\left(\frac{Ln+1}{Kn+1}\right)\sigma_{gI}^2} \quad . \tag{A2}$$

In comparing Equations A1 and A2, it is evident that $(E\rho_g^2 \mid P) < E\rho_p^2$ when

$$\frac{L+1}{K+1} < \frac{Ln+1}{Kn+1} \quad ,$$

which is equivalent to

$$(L + 1)(Kn + 1) < (K + 1)(Ln + 1)$$

$$LKn + L + Kn + 1 < LKn + K + Ln + 1$$

$$Kn - K < Ln - L$$

$$K(n - 1) < L(n - 1)$$

$$K < L$$

Therefore, $E\rho_g^2 | P < E\rho_p^2$ when

$$\sigma_g^2 / \sigma_{p:g}^2 < \sigma_{gl}^2 / \sigma_{pl:g}^2$$

which is equivalent to

$$\frac{\sigma_g^2}{\sigma_g^2 + \sigma_{p:g}^2} < \frac{\sigma_{gi}^2}{\sigma_{gi}^2 + \sigma_{pi:g}^2} \quad ,$$

as shown in the text leading to Equation 11.